ABSTRACT

One embodiment is a method to identify a preprocessing algorithm for raw data. The method may includes the steps of providing an algorithm knowledge database including preprocessing algorithm data and feature set data associated with the preprocessing algorithm data, analyzing raw data to produce analyzed data, extracting from the analyzed data features that characterize the data, and selecting a preprocessing algorithm using the algorithm knowledge database and features extracted from the analyzed data. Another embodiment is a data mining system for identifying a preprocessing algorithm for raw data using this method. Still another embodiment is a data mining application with improved preprocessing algorithm selection, including (a) an algorithm knowledge database containing preprocessing algorithm data and feature set data associated with the preprocessing algorithm data; (b) a data analysis module adapted to receive control of the data mining application when the data mining application begins; (c) a feature extraction module adapted to receive control of the data mining application from the data analysis module and available to identify a set of features; and (d) an algorithm selection module available to receive control from the feature extraction module and available to identify a preprocessing algorithm based upon the set of features identified by the feature extraction module using the algorithm knowledge database.